Computers and Artificial Intelligence；Vol.3, No.3; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n3.193199
Published by: Zeus Press

Reliability Evaluation of Large Language Models for Social Media Sentiment Annotation: An Empirical Study Based on Model Agreement and Downstream Tasks

Wenjing Pi[†] and Changxian He[†,*]

Dongbei University of Finance and Economics, Dalian, Liaoning, China

[†]These two authors contributed equally to this work.

**Corresponding author: Changxian He.*

Abstract

Driven by booming social media and user-generated texts, sentiment analysis stands as a core natural language processing task, yet large-scale high-quality data labeling comes with steep costs and practical barriers. This work assesses how dependable large language models are for sentiment tagging, alongside how their labeled outputs shape subsequent classification effects. We build a dataset containing 1,543 Chinese entertainment comment snippets scraped from Bilibili. Under unified prompting, three LLMs—DeepSeek, Qwen and Doubao—produce zero-shot sentiment tags, while 500 sampled entries receive manual annotation to form authoritative benchmark labels. Cohen's Kappa is adopted to quantify human-model annotation consistency, and TF-IDF features are input to Logistic Regression, Linear SVM and Random Forest for downstream classification evaluation. Among the three models, Doubao achieves the highest human-label consistency with a κ value of 0.800, exceeding Qwen (κ=0.746) and DeepSeek (κ=0.636). Classifiers trained on Doubao's labels obtain optimal Macro-F1 values of 0.8125 (SVM) and 0.8174 (Random Forest). Obvious performance discrepancies exist across LLMs; high-quality annotations significantly boost downstream classification accuracy, which highlights the importance of selecting competent LLMs for sentiment labeling tasks.

Keywords

large language models (LLMS), sentiment analysis, user-generated content, zero-shot learning

1. Introduction

The rapid development of social media platforms has spawned massive user-generated textual content, typical of which are comment texts on Bilibili [1]. Containing rich subjective emotional information, such social media texts have become key research objects of sentiment analysis, and are widely applied to public opinion monitoring, user behavior analysis and personalized recommendation systems [2, 3]. At present, mainstream supervised sentiment analysis methods rely heavily on manually annotated corpora for model training [4]. However, manual annotation suffers from high labor costs, time-consuming processes and low scalability, which greatly limits its application in large-scale real-world scenarios. With the remarkable advancement in zero-shot learning, large language models (LLMs) have been widely applied to text

classification and automatic annotation, emerging as an effective alternative to manual labeling [5]. Traditional sentiment classification methods that combine TF-IDF feature extraction with classic machine learning algorithms, including Logistic Regression and Support Vector Machines, are still prevalent due to their simplicity and interpretability. However, traditional feature-based methods are unable to mine profound semantic and contextual features within texts, as illustrated by previous research [6]. methods built upon large language models possess prominent advantages in zero-shot sentiment classification tasks, which is why such approaches have been widely adopted to realize automated text annotation within existing relevant studies [7]. While LLMs are commonly used to generate annotation labels, limited existing work has quantitatively verified the stability and trustworthiness of such machine-produced labels, leading to inadequate empirical proof to confirm annotation reliability [8]. Most prior literature fails to carry out parallel comparative analysis among mainstream LLMs for sentiment labeling tasks. Differences between model-generated annotations and human gold-standard labels remain insufficiently explored, leaving the practical performance of various LLM annotation pipelines unclear. Hence, this study constructs an integrated evaluation system to comprehensively measure the quality and consistency of annotations from different LLMs.

According to relevant statistical research, the Cohen's Kappa coefficient acts as a standard and extensively used statistical metric to measure consistency between multiple annotators in classification experiments [9]. It excludes the interference of random label matching and can objectively and quantitatively measure the annotation consistency between human annotators and models, serving as a standard evaluation metric for annotation quality assessment.As typical Chinese social media text, Bilibili entertainment comments are characterized by short sentences, colloquial expressions, internet slang and implicit sentiment expressions, which pose higher requirements for model semantic understanding and sentiment discrimination capability. Accordingly, this study constructs a multi-LLM comparative evaluation framework to explore three core research issues: the consistency of annotation results among different LLMs, the agreement between LLM-generated labels and human gold standards, and the influence of annotation quality on downstream classification performance. This study adopts Cohen's Kappa coefficient to quantify inter-model and human-model annotation consistency, and conducts downstream classification experiments based on TF-IDF features and traditional machine learning models to reveal the inherent correlation between annotation quality and sentiment classification performance.

2. Materials and Methods

2.1 Dataset Collection and Preprocessing

This study uses entertainment-related comment data from Bilibili. Raw data were collected automatically via Shadowbot RPA, followed by data cleaning to eliminate blank content, emoji-only texts and punctuation-only sentences, leaving 1,543 valid samples in the final dataset. A three-way sentiment taxonomy (Positive, Neutral, Negative) was adopted, numerically coded as 1, 2 and 3 separately. All valid texts were preprocessed and matched in Excel to remove redundant spaces and other noise before annotation. After all samples were annotated by LLMs, 500 entries were randomly sampled as a separate test set for manual labeling. These human-labeled samples were excluded from model training and prompt construction to guarantee objective and unbiased evaluation.

2.2 LLM Annotation Strategy

Considering that the collected comments were written in Chinese, three mainstream Chinese Large Language Models (LLMs), namely DeepSeek, Qwen, and Doubao, were selected for sentiment annotation. Specifically, DeepSeek adopts the API model ID deepseek-chat, Qwen uses qwen-plus, and Doubao corresponds to doubao-seed-character-251128. To ensure experimental consistency, all models were prompted using an identical Chinese Zero-shot prompt, and each comment was assigned a single sentiment label. This setting minimized potential confounding factors and enabled a fair comparison among different models. The specific model configurations are summarized in Table 1.

Table 1: Large Language Models and API Configurations

Model	API Model ID
DeepSeek	deepseek-chat

Qwen	qwen-plus
Doubao	doubao-seed-character-251128

2.3 Inter-Model Agreement Evaluation

To quantify annotation consistency disparities among the three LLMs, Cohen's Kappa coefficient was used for pairwise comparison of labeling results. All statistical analyses were carried out in SPSS 23.0, with Kappa value interpretation criteria adopted from established existing research [10]. In detail, κ values equal to or lower than zero had signified a complete absence of inter-model annotation consistency; values ranging from 0.01 to 0.20 had corresponded to only slight agreement among model labeling results; values between 0.21 and 0.40 had represented a fair level of annotation consistency; values falling within the interval of 0.41 to 0.60 had indicated moderate inter-model alignment; values from 0.61 to 0.80 had demonstrated substantial consistency across different LLM annotations, while κ scores fluctuating between 0.81 and 1.00 had denoted an almost identical and near-perfect annotation performance among the tested large language models.

2.4 Construction of the Human-annotated Gold Standard

A trustworthy benchmark corpus was built for subsequent quality validation. Five hundred text entries were randomly sampled out of the full dataset, and separate labeling work was finished independently by two human coders. Before these manual labels could act as authoritative ground truth, the Cohen's Kappa coefficient measuring alignment between the two annotators was computed to verify the stability of human tagging outcomes. The initial independent annotation yielded $\kappa=0.939$. Per widely accepted criteria [10], this value falls within 0.81–1.00, representing near-perfect inter-annotator agreement and verifying consistent labeling standards as well as dataset reliability.

For samples with inconsistent labels, two annotators discussed discrepancies to reach unified sentiment categories. These consensus labels served as the human gold standard for evaluating LLM annotation quality and downstream classification performance.

2.5 Downstream Classification Tasks

The core objective of this experimental module lies in revealing how annotation quality of different degrees would exert differentiated influences on the predictive performance exhibited by subsequent sentiment classification models. To transform raw comment texts into computable numerical feature vectors, the TF-IDF feature extraction mechanism that integrates both unigram and bigram lexical units had been adopted throughout the whole process. Three mainstream traditional machine learning approaches had been chosen to establish separate classification pipelines in this research, including Logistic Regression (LR), Linear Support Vector Machine (Linear SVM) as well as Random Forest (RF). For each group of comparative experiments, sentiment labels automatically generated by a single large language model would serve as the training dataset. In contrast, all 500 manually annotated gold-standard samples had been retained in advance and would be consistently adopted as the fixed test set across every model comparison group.

2.6 Evaluation Metrics

In order to conduct a comprehensive and unbiased quantitative assessment on the classification capacity delivered by all experimental machine learning models that had been built within this research framework, the Macro-F1 metric had been designated as the core and priority evaluation indicator throughout the whole series of downstream sentiment classification tests. This specialized statistical indicator was capable of producing balanced and fair evaluation outcomes covering every pre-defined sentiment category, owing to its inherent mechanism that could eliminate biased prediction errors triggered by unbalanced sample quantities belonging to majority emotional labels. Beyond this core measurement standard, the global classification accuracy rate had also been introduced as an auxiliary evaluation benchmark, through which researchers could directly observe and quantify the universal generalization performance as well as practical application value possessed by each separately constructed machine learning classifier.

3. Results and Analysis

3.1 Inter-LLM Agreement Results

The pairwise annotation agreement among the three Large Language Models is presented in Table 2. Overall, all model pairs exhibited at least moderate agreement, indicating that the three LLMs shared a relatively consistent understanding of sentiment categories in social media texts.

Table 2: Inter-LLM Agreement Results

Model	Cohen's κ
DeepSeek vs Qwen	0.683
Qwen vs Doubao	0.664
DeepSeek vs Doubao	0.586

Among the three model pairs, the highest agreement was observed between DeepSeek and Qwen, with a Cohen's Kappa coefficient of 0.683. The agreement between Qwen and Doubao was slightly lower ($\kappa = 0.664$), whereas the agreement between DeepSeek and Doubao was comparatively weaker ($\kappa = 0.586$). Although differences existed among the model pairs, all κ values exceeded 0.50, suggesting that the sentiment annotations generated by different LLMs maintained a moderate-to-substantial level of consistency.

3.2 Human–LLM Agreement Results

The agreement between the three Large Language Models and human annotations is summarized in Table 3.

Table 3: Human–LLM Agreement Results

Model	Cohen's κ
DeepSeek vs human	0.636
Qwen vs human	0.746
Doubao vs human	0.800

Experimental results demonstrated that the Doubao model attained the highest consistency with manual annotation results, which indicated that its sentiment discrimination decisions were most consistent with the established human-annotated gold standard. Among the three evaluated LLMs, Qwen achieved a moderate level of annotation agreement and occupied the second position, while DeepSeek obtained the lowest human-model consistency throughout the annotation task.

When the three predefined sentiment categories were further analyzed against the manually calibrated gold standard dataset, the results had demonstrated that neutral comment texts had constituted the most intractable annotation challenge for all three adopted large language models. While positive and negative user comments had conveyed explicit emotional inclinations that were easy to identify, neutral textual content had inherently embodied ambiguous semantic implications and indistinct categorical demarcations, which had inevitably elevated the misclassification probability and compromised the annotation accuracy of all tested LLMs. In comparison, the overt emotional expression embedded within positive and negative social media comments had allowed all models to produce more stable and consistent annotation outcomes on these two sentiment categories.

Among the three competing large language models that had participated in the zero-shot annotation task, Doubao had possessed outstanding comprehensive competence in coping with texts carrying uncertain and blurred semantic features. This distinctive technical advantage had effectively minimized its misjudgment frequency on ambiguous neutral samples, which had ultimately contributed to its superior overall consistency with authoritative human annotation standards.

3.3 Downstream Classification Performance

The comprehensive classification performance of diverse downstream machine learning models that had been trained on pseudo-labels generated by different large language models had been systematically organized and displayed in Table 4.

Table 4: Downstream Classification Performance

Model	LR F1	SVM F1	RF F1
DeepSeek	0.3916	0.6078	0.6097
Qwen	0.3242	0.7367	0.7402
Doubao	0.3770	0.8125	0.8174

As reflected by the quantitative experimental outcomes, all three conventional machine learning classifiers had attained their optimal predictive performance when the training datasets were constructed based on the annotation results output by the Doubao model. Within the Logistic Regression classification framework, noticeable performance discrepancies had been observed among the three LLMs in terms of Macro-F1 values. DeepSeek had achieved a Macro-F1 score of 0.3916 and Doubao had attained 0.3770, whereas Qwen had only obtained a score of 0.3242, which had placed it at the bottom of the performance ranking for the LR experiment. In sharp contrast, Linear SVM and Random Forest models had presented far more distinguishable performance disparities when trained on labels from different model sources. The Linear SVM classifier had yielded Macro-F1 values of 0.6078, 0.7367 and 0.8125 using annotation data from DeepSeek, Qwen and Doubao respectively, while the Random Forest algorithm had generated parallel results of 0.6097, 0.7402 and 0.8174 under identical experimental configurations. Such results had clearly proven that the high-quality annotation labels provided by Doubao had brought remarkable performance gains to these two downstream classification models.

The aforementioned experimental phenomena had solidly verified the positive facilitating effect of high-precision LLM automatic annotation on downstream model training. Those large language models that had maintained higher annotation consistency with human gold standards were capable of producing more reliable pseudo-label datasets, which had further strengthened the generalization ability and overall predictive efficacy of subsequent machine learning classifiers. Furthermore, the performance hierarchy of all experimental groups in downstream classification tasks had precisely matched the ranking of human-model annotation consistency, which had powerfully confirmed the existence of a stable positive linear correlation between the reliability of LLM-generated labels and the final classification performance of traditional machine learning algorithms.

4. Discussion

This study conducts quantitative comparisons of reliability differences among multiple LLMs via a three-class sentiment annotation task using Bilibili entertainment comments. The evaluation framework integrates zero-shot annotation results, human gold-standard comparison and downstream machine learning verification. Combined with the linguistic features of Chinese social media texts and weakly supervised learning theory, this paper analyzes the mechanisms accounting for experimental results.

4.1 Underlying Factors Affecting Annotation Consistency

Comparisons between model outputs and human labels reveal obvious gaps in annotation performance. Bilibili comments consist of short informal sentences, internet slang and implicit sentiment, which place high demands on models' semantic understanding and sentiment recognition [11]. Prior studies indicate inconsistent training data, model architectures and instruction-following abilities are core contributors to divergent annotation results on complex sentiment tasks [12].

The three Chinese general LLMs adopted in this work differ greatly in training objectives. DeepSeek is optimized for general knowledge and dialogue QA, making it poorly suited to fragmented colloquial social media text. Qwen and Doubao are specially tuned for conversational online content, yielding more stable sentiment labels.

The zero-shot setting eliminates annotated examples that could standardize classification logic, widening performance gaps across LLMs. It demonstrates that LLM annotation effectiveness is highly task- and domain-dependent.

4.2 Propagation of Annotation Noise to Downstream Tasks

When interpreted under the theoretical framework of weakly supervised learning, all sentiment labels produced by large language models could be regarded as a type of pseudo-labels in model training. Throughout the entire experimental procedure, the Cohen's Kappa coefficient calculated between human annotators and

each LLM had been utilized to quantitatively reflect the severity of noise contained within machine-generated annotations. All recorded experimental observations had validated the inherent rule that a weaker consistency degree between human benchmarks and model outputs would correspond to more severe label noise, which would further degrade the predictive performance of downstream classifiers; such a discovery could align perfectly with previous research conclusions pointing out that low-quality pseudo-labels would create negative interference during the fitting phase of machine learning models [13].

Among the three selected large language models, Doubao had been capable of generating the most reliable annotation results with clear dividing lines between distinct sentiment categories. When Linear SVM and Random Forest algorithms were trained upon these high-quality labels, the two classifiers could reach Macro-F1 values of 0.8125 and 0.8174 separately. In contrast, DeepSeek presented relatively low consistency against human labeling standards, and the vague, conflicting labels it produced would eventually lead to sharply decreased classification performance, with its corresponding Macro-F1 metrics dropping to merely 0.6078 and 0.6097.

Logistic Regression had displayed inferior classification performance within every experimental subgroup. Compared with regularized SVM and ensemble learning frameworks, this algorithm would be far more susceptible to the adverse impacts brought by noisy labels. Furthermore, short online comment texts would create highly sparse feature vectors after TF-IDF transformation, which weakened the discriminative power of feature information and put extra constraints on the predictive capacity of Logistic Regression. The completely matched rank order between human-model consistency levels and downstream classification performance had proven the existence of a positive correlation linking annotation quality and the final effectiveness of sentiment classification.

4.3 Limitations and Future Work

Three obvious drawbacks had existed within the current research, which would restrict the external generalization ability of all experimental conclusions. To start with, only entertainment comment data collected from Bilibili had been adopted for all experimental tests, so the research findings could not be reasonably extended to text datasets sourced from other domains or social media platforms. Next, the study had only examined the zero-shot annotation paradigm, without conducting supplementary exploration on alternative prompt strategies such as few-shot learning and chain-of-thought reasoning. Last but not least, traditional machine learning classifiers had served as the sole downstream evaluation tools, while deep learning architectures and transformer-based pre-trained models were not incorporated into the comparison experiments.

Multiple expansion avenues can be pursued in follow-up investigations. Diverse datasets covering varied platforms and domains may be integrated to strengthen result robustness. Extra mainstream LLMs and assorted prompt strategies can be compared to unpack how prompt engineering reshapes labeling quality. Pre-trained language models can also serve as downstream baselines to test deep learning's tolerance to labeling noise. Beyond that, multimodal large language models deployed for zero-shot sentiment tagging constitute a promising research branch waiting for in-depth exploration [14].

5. Conclusions

This paper carries out comprehensive assessments on annotation consistency and reliability of three LLMs using Bilibili social media comments. Obvious gaps in labeling quality are captured among tested models, with Doubao delivering the most robust annotation performance.

A positive correlation is verified between label quality and downstream classification accuracy. Models whose tags match human gold standards more closely supply higher-quality training signals and yield superior classification results. These discoveries verify that high-fidelity LLM-generated labels effectively boost the overall performance of subsequent natural language processing tasks.

Collectively, this research furnishes empirical evidence for reliability evaluation of LLMs in automatic sentiment labeling, and delivers actionable references for model screening and corpus construction targeting social media textual analysis.

References

[1] Li X, Li J, Wu Y. A global optimization approach to multi-polarity sentiment analysis. PLoS One. 2015 Apr 24;10(4):e0124672.

[2] Zhu L, Xu M, Bao Y, Xu Y, Kong X. Deep learning for aspect-based sentiment analysis: a review. PeerJ Comput Sci. 2022 Jul 19;8:e1044.

[3] Ruan T, Liu Q, Chang Y. Digital media recommendation system design based on user behavior analysis and emotional feature extraction. PLoS One. 2025 May 19;20(5):e0322768.

[4] Bilal M, Khan A, Jan S, Musa S, Ali S. Roman Urdu Hate Speech Detection Using Transformer-Based Model for Cyber Security Applications. Sensors (Basel). 2023 Apr 12;23(8):3909.

[5] El Koshiry AM, Eliwa EHI, Abd El-Hafeez T, Khairy M. Detecting cyberbullying using deep learning techniques: a pre-trained glove and focal loss technique. PeerJ Comput Sci. 2024 Mar 27;10:e1961.

[6] Jin T, Liu J. A text classification method by integrating mobile inverted residual bottleneck convolution networks and capsule networks with adaptive feature channels. Sci Rep. 2025 Jan 5;15(1):855.

[7] Luo, Han, Cai, Meng, Cui, Ying, Spread of Misinformation in Social Networks: Analysis Based on Weibo Tweets, Security and Communication Networks, 2021, 7999760, 23 pages, 2021.

[8] Satheakeerthy S, Jesudason D, Bahrami B, Bacchi S, Lee YM, Casson R, Sun M, Chan W. Zero-shot LLM-based visual acuity extraction: a pilot study. BMC Ophthalmol. 2025 Jul 1;25(1):359.

[9] Chen J, Zhang X, Wang J, Cao T, Gu C, Li Z, Song Y, Yang L, Zhang Z, Zhang Q, Qian D, Li X. Renji endoscopic submucosal dissection video data set for colorectal neoplastic lesions. Sci Data. 2025 Aug 6;12(1):1366.

[10] McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 22(3):276–282

[11] Zhou K, Chen W, Sheng W, Hu B, Yang Y, Yang L, Lu H. Explainable AI with fine-tuned large language models for sustainable cultural heritage management. Sci Rep. 2025 Nov 21;15(1):41370.

[12] Garcia-Carmona AM, Prieto ML, Puertas E, Beunza JJ. Leveraging Large Language Models for Accurate Retrieval of Patient Information From Medical Reports: Systematic Evaluation Study. JMIR AI. 2025 Jul 3;4:e68776.

[13] Huang Y, Lu J, Wu Q, Zhang G. Weakly Supervised Composed Object Re-Identification With Large Models. IEEE Trans Cybern. 2026 Apr 17;PP.

[14] Muhammad Umair Ali et al. From vectors to knowledge graphs: A comprehensive analysis of modern retrieval-augmented generation architectures. Computer Science Review. 2026;61:100925.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights